Data Mining for Anomaly Detection from Numeric and Text Data

University of Minnesota
United Technologies Research Center

NASA Aviation Safety Technical Conference
October 21 – 23, 2008

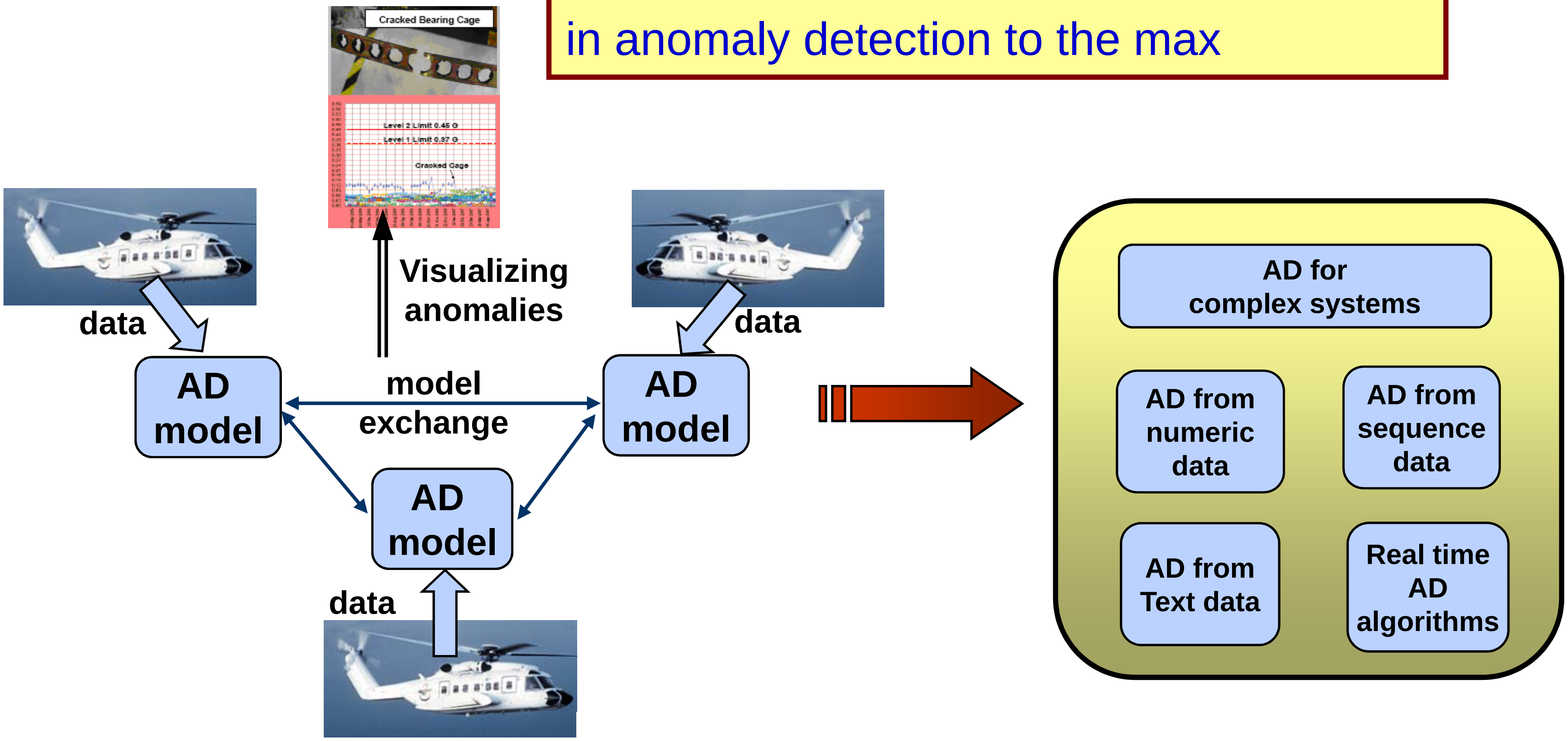
Project Team

- **University of Minnesota**
  - Jaideep Srivastava, PI
  - Vipin Kumar, Co-PI
  - William Schuler, Co-PI
  - Arindam Banerjee, Co-PI
  - Students & Researchers
    - **Varun Chandola, Nishith Pathak, Hanhuai Shan, Junlin Zhou, Mingsheng Shang, Tim Miller, Lane Scwartz, Stephen Wu**
- **UTRC**
  - Aleksander Lazarevic, Co-PI

The Big Picture

Research Objective: Detect anomalous events & trends from multiple, heterogeneous, distributed data sources for complex systems, in real time

In other words: Push the state-of-the-art in anomaly detection to the max



Anomaly detection from data with mixed continuous and discrete attributes

Anomaly Detection for Continuous Sequences

Problem Statement

- Given a set of test sequences and a set of normal training sequences, assign an anomaly score to each test sequence with respect to the training set.
  - Sequences are univariate continuous (or univariate time-series).
  - Sequences can be of variable lengths.

- Developed a library (SQUAD) of anomaly detection techniques for symbolic sequences.
  - Allows using six different techniques for anomaly detection.
  - Allows using six different methods to combine per event probabilities into a combined anomaly score for the test sequence.
  - Written in C,C++, and Perl.

Results

	motor1	motor2	motor3	motor4	valve1	power	chf01	chf02	ltstdb21	ltstdb31	mitdb06	mitdb19	edb03	edb05
Euclid.	0.70	0.70	0.70	0.90	1.00	0.62	0.12	0.12	0.10	0.12	0.14	0.08	0.22	0.26
DTW	0.80	0.90	0.70	1.00	1.00	0.88	0.18	0.64	0.46	0.20	0.84	0.84	0.80	0.22
SMC	0.70	0.50	0.70	0.50	0.88	0.88	0.14	0.16	0.14	0.28	0.46	0.48	0.60	0.12
wSMC	0.70	0.70	0.70	0.80	0.75	0.75	0.12	0.16	0.10	0.16	0.52	0.66	0.74	0.22
nLCS	1.00	0.90	1.00	0.90	0.88	0.88	0.08	0.20	0.14	0.26	0.42	0.46	0.62	0.16
DISCORD (Cont.)	1.00	1.00	1.00	1.00	0.88	0.75	0.24	0.68	0.64	0.66	0.58	0.74	0.76	0.26
DISCORD (Disc.)	0.50	0.50	0.50	0.50	0.75	0.88	0.12	0.12	0.24	0.28	0.48	0.58	0.76	0.18
tSTIDE	0.70	0.70	0.80	0.80	1.00	0.62	0.18	0.26	0.16	0.26	0.36	0.48	0.42	0.18
SVR	1.00	1.00	1.00	1.00	0.75	0.75	0.04	0.08	0.04	0.08	0.24	0.88	0.90	0.30
FSAz	0.80	0.70	0.80	0.80	1.00	0.75	0.18	0.26	0.10	0.26	0.38	0.76	0.36	0.18

Conclusions

MMultiple techniques can be applied to detect anomalies in continuous sequences.

PPerformance of various techniques depends on the nature of the underlying data.

((SAX) Discretization based techniques perform poorly compared to their continuous counterparts.

KKNN based technique using DTW, DISCORD, and SVR are the most consistent techniques.

PPerformance of kNN and SVR is better when the anomalous and normal sequences are generated from a different source.

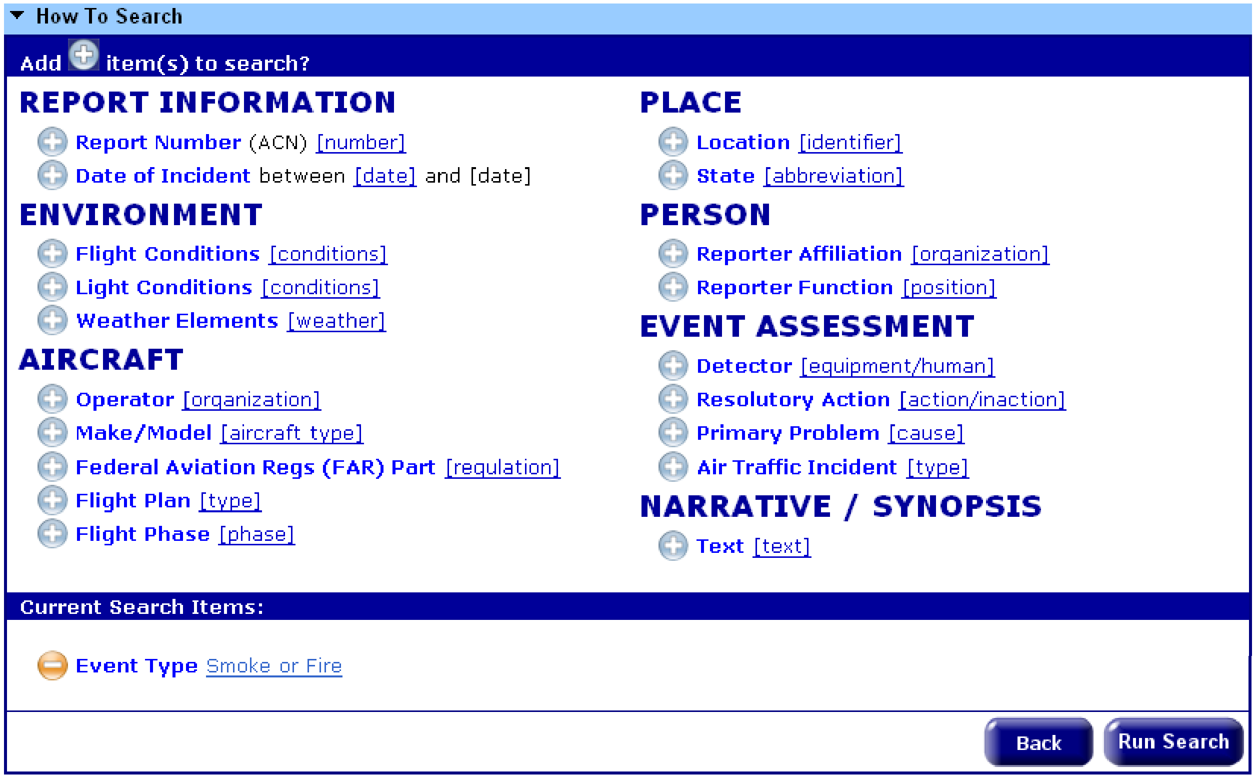
DDISCORD technique is well suited for the case when the anomalous sequences are minor deviations of the normal sequences.

Anomaly Detection from Databases of Textual Reports

United Technologies

Research Center

ASRS Database



▼ How To Search

Add ⊕ item(s) to search?

REPORT INFORMATION

- ⊕ Report Number (ACN) [number]
- ⊕ Date of Incident between [date] and [date]

ENVIRONMENT

- ⊕ Flight Conditions [conditions]
- ⊕ Light Conditions [conditions]
- ⊕ Weather Elements [weather]

AIRCRAFT

- ⊕ Operator [organization]
- ⊕ Make/Model [aircraft type]
- ⊕ Federal Aviation Regs (FAR) Part [regulation]
- ⊕ Flight Plan [type]
- ⊕ Flight Phase [phase]

PLACE

- ⊕ Location [identifier]
- ⊕ State [abbreviation]

PERSON

- ⊕ Reporter Affiliation [organization]
- ⊕ Reporter Function [position]

EVENT ASSESSMENT

- ⊕ Detector [equipment/human]
- ⊕ Resolutory Action [action/inaction]
- ⊕ Primary Problem [cause]
- ⊕ Air Traffic Incident [type]

NARRATIVE / SYNOPSIS

- ⊕ Text [text]

Current Search Items:

⊖ Event Type Smoke or Fire

Narratives report an anomaly:

I WAS FLYING THE KATANA WITH A STUDENT AND ON DOWNWIND **THE FUEL PRESSURE DROPPED TO ZERO, AND THE ENG WAS CUTTING OFF**. I VERIFIED FUEL PUMP WAS ON AND IT WAS ON. BY THE TIME WE TURNED SHORT FINAL, THE PROP STOPPED AND WE LANDED THE AIRPLANE SAFELY. THEN WE CALLED CASTLE UNICOM TO SEND THE FUEL TRUCK

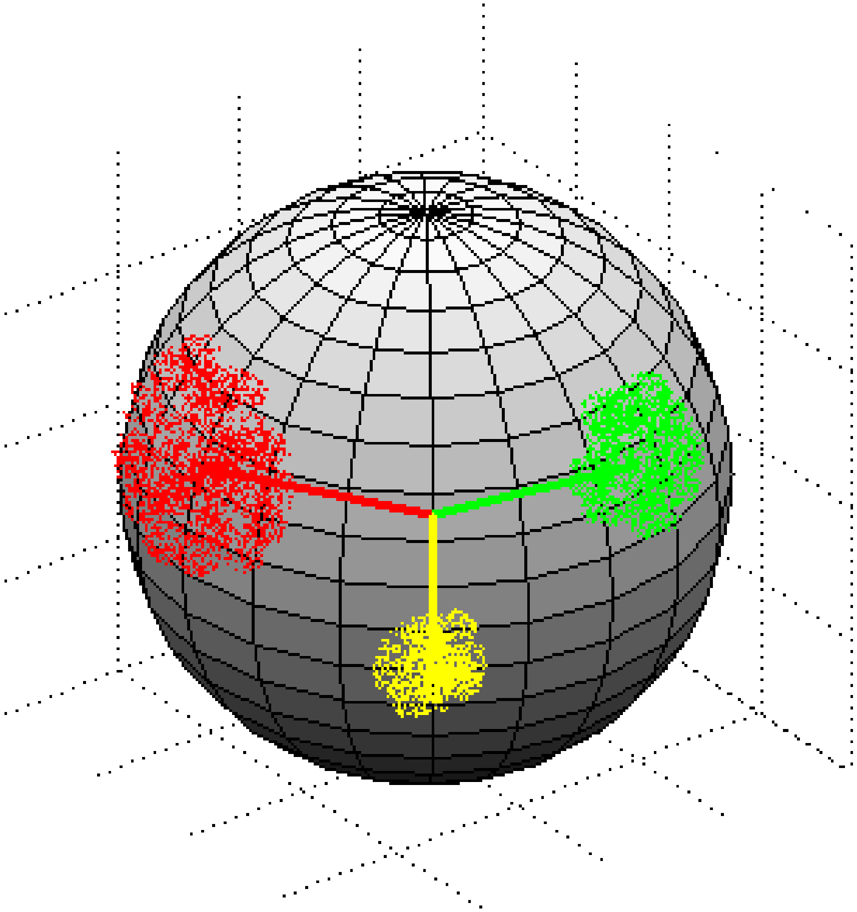
Goal

- Automatically discover various types (categories) of anomalies from textual reports.
  - e.g. Maintenance, Weather…
  - Why?

> …RPTR FURTHER STATED THAT **THIS HAS BEEN A PROBLEM FOR SEVERAL YEARS WITH VERY LITTLE DONE BY THE ARPT**…

- Put each report into a certain category/categories.
  - Which report addresses which problem(s).
  - Correct the reports that are in wrong categories in the database.

Mixture of von Mises Fisher Distribution [Banerjee *et al,* 2005]



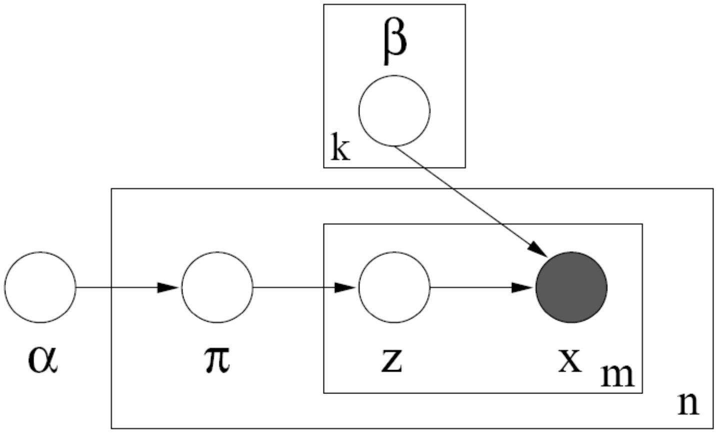
•Data points (reports) lie on a unit hyper-sphere.

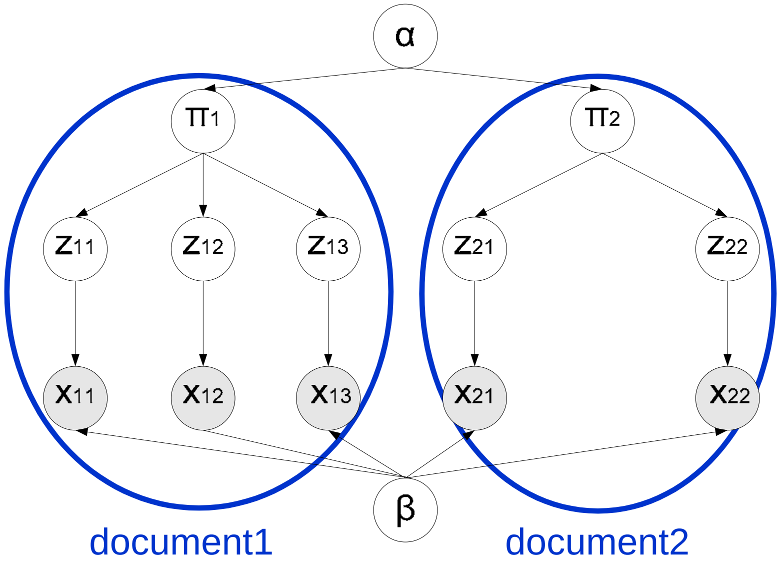
- *mean direction*

- *concentration parameter*

• Example: Three types of reports could be represented by three vMF distributions (red, green, yellow) – mixture of vMF.

Latent Dirichlet Allocation [Blei *et al*, 2003]





- For each document,
- Choose π ~Dirichlet(α)
- For each word x_n:
  - Choose a topic z_n ~Discrete(π)
  - Choose a word x_n from $p(x_n|z_n,\beta)$, a Discrete distribution conditioned on the topic z_n.

Confusion Matrix and Topic Lists for a Three-category Dataset

Dataset: *NASA* - 4226 Reports, three causes of the problem

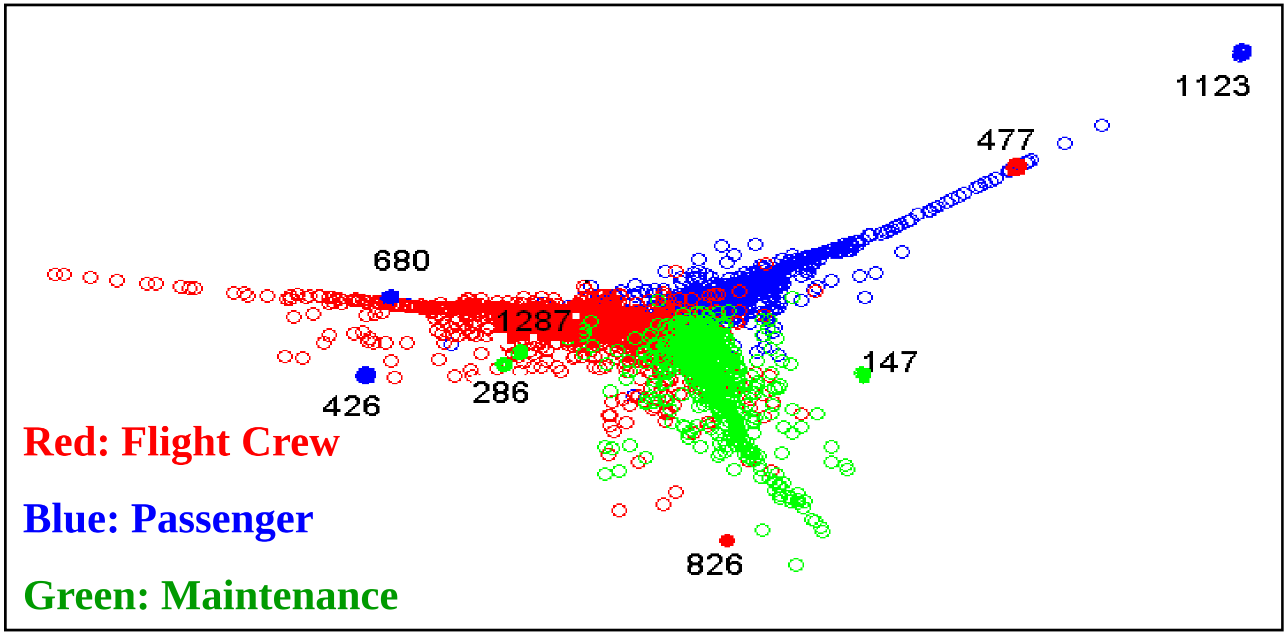
- Flight crew human performance.
- Passenger.
- Maintenance human performance

	1	*2*	*3*
1	**1185**	45	35
2	12	**1150**	49
3	169	42	**1538**

Numbers on the diagonal –number of correctly clustered reports

Flight Crew	Passenger	Maintenance
rwy	pax	acft
apch	flt	maint
acft	attendant	eng
dep	capt	zzz
alt	seat	flt
turn	told	mel
time	asked	chk
atc	back	fuel
flt	attendants	time
twr	acft	gear

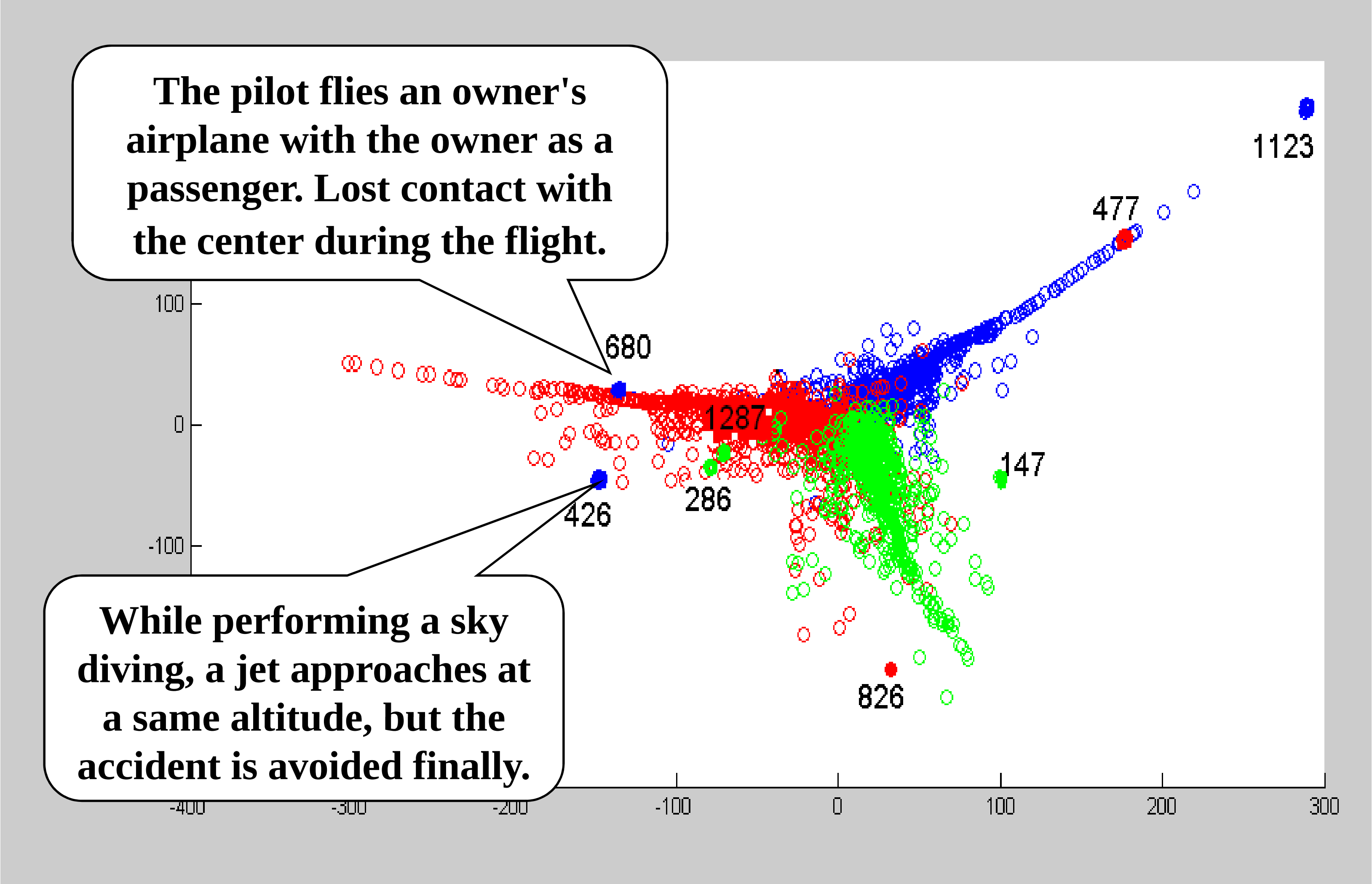
Two-Dimensional Visualization for Reports



- Each point represents one report.
- The color of the point – the report's label.
- The location of the point – mixed membership from LDA +ISOMAP.
- Focusing on: Points having different colors with the neighbours

  Isolated points

Two-Dimensional Visualization for Reports

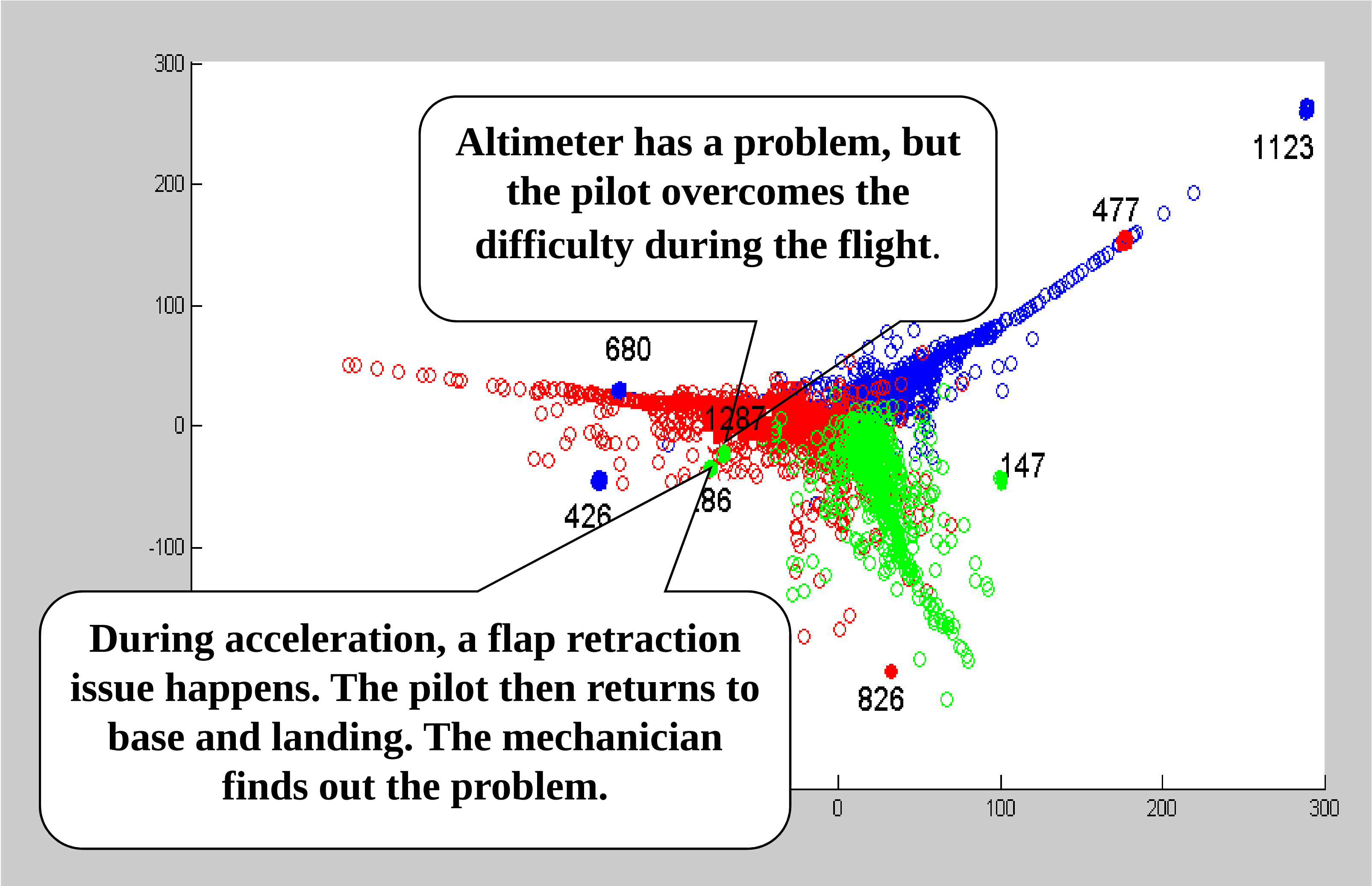


The pilot flies an owner's airplane with the owner as a passenger. Lost contact with the center during the flight.

While performing a sky diving, a jet approaches at a same altitude, but the accident is avoided finally.

Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports

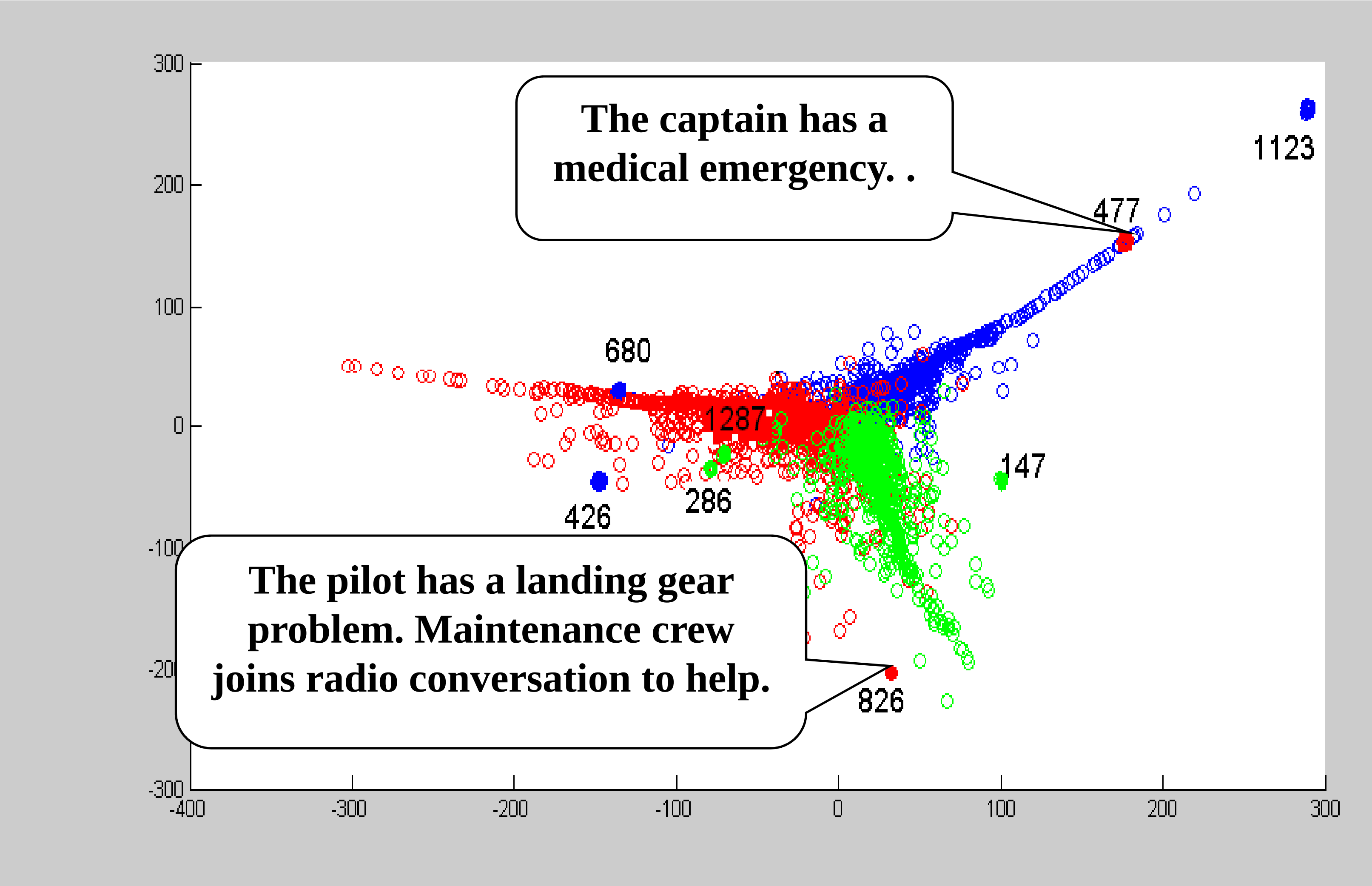


Altimeter has a problem, but the pilot overcomes the difficulty during the flight.

During acceleration, a flap retraction issue happens. The pilot then returns to base and landing. The mechanician finds out the problem.

Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports

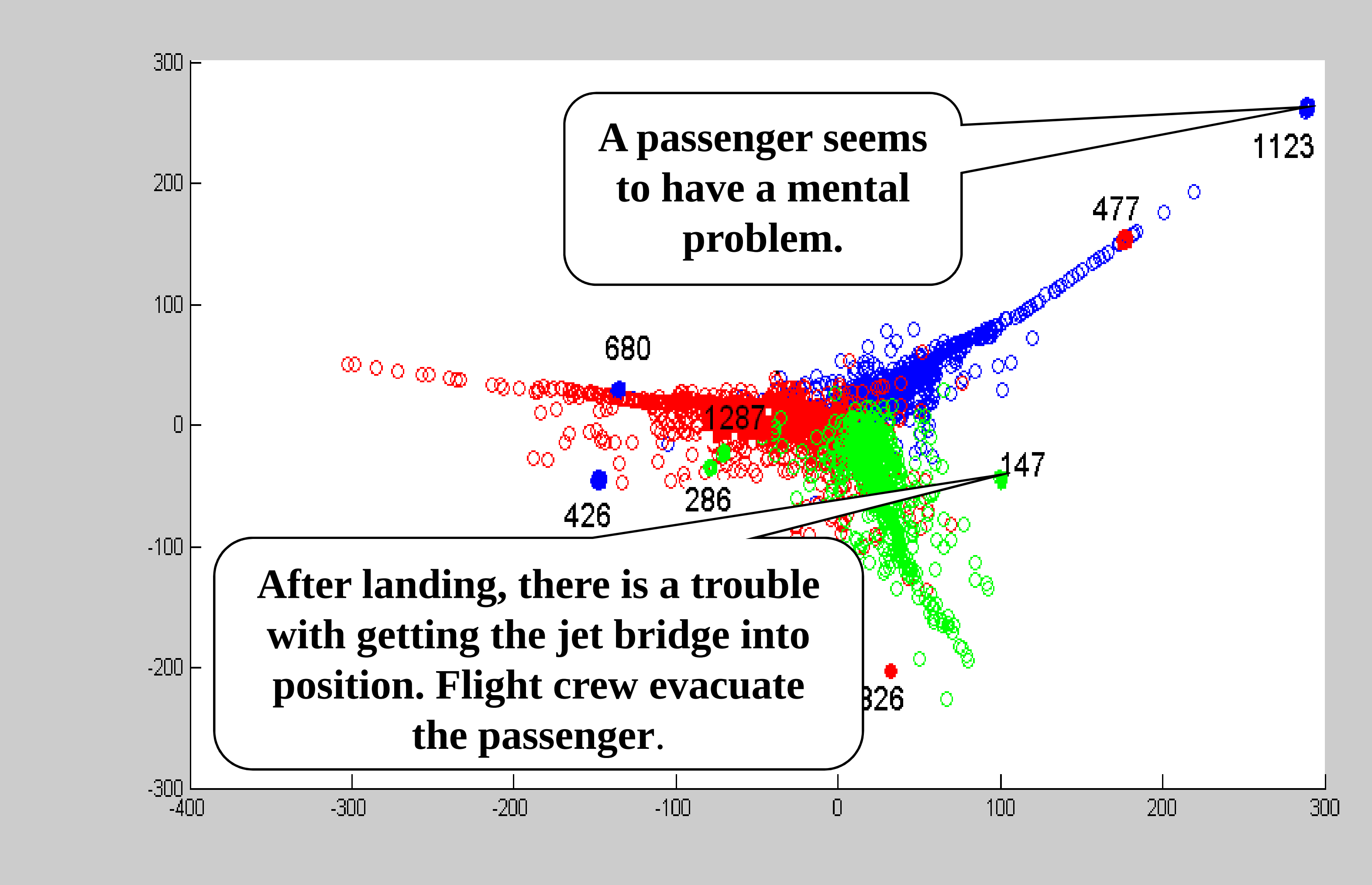


The captain has a medical emergency. .

The pilot has a landing gear problem. Maintenance crew joins radio conversation to help.

Red: Flight crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



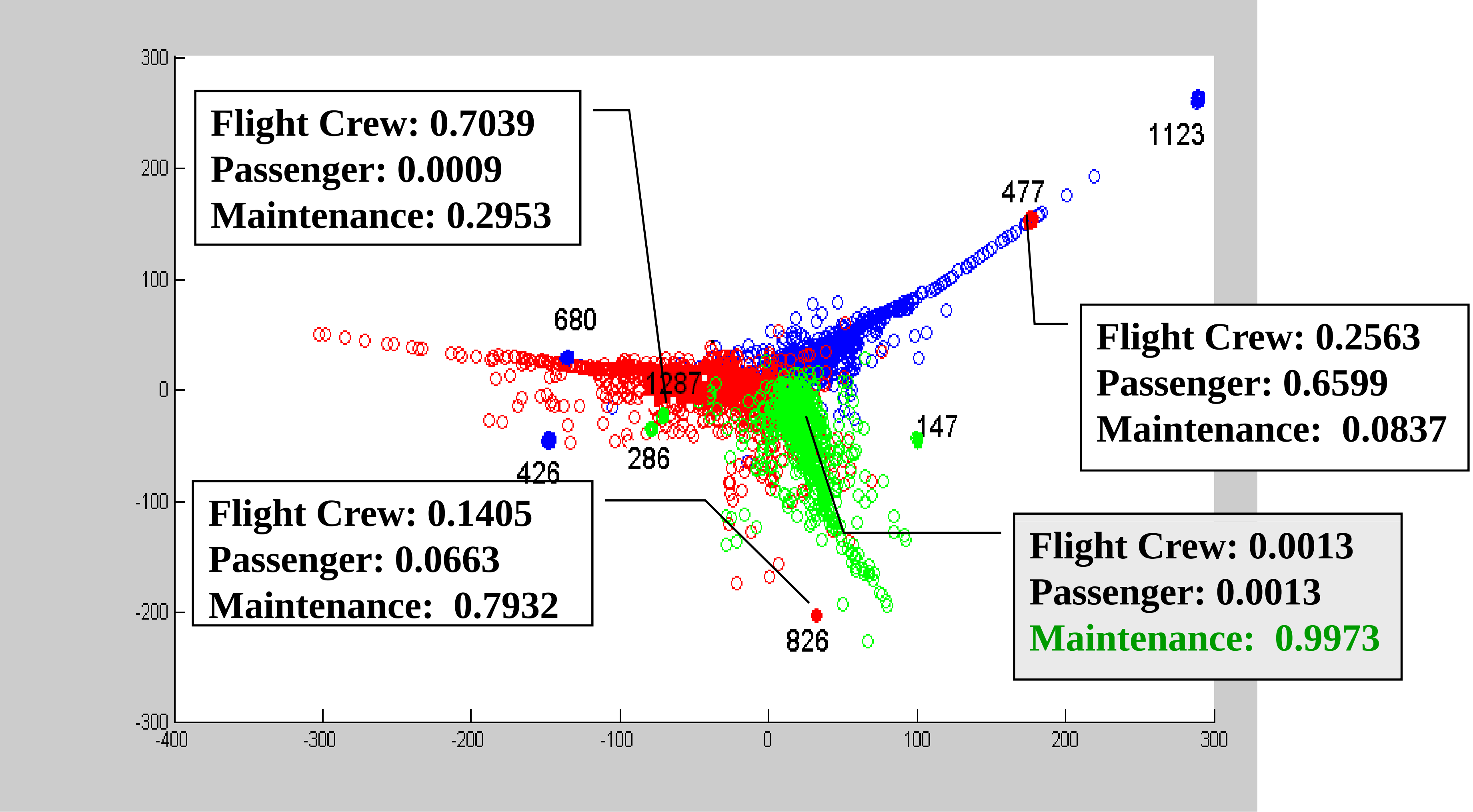
A passenger seems to have a mental problem.

After landing, there is a trouble with getting the jet bridge into position. Flight crew evacuate the passenger.

Red: Flight Crew Blue: Passenger Green: Maintenance

United Technologies Research Center

Mixed Membership of Reports



Flight Crew: 0.7039
Passenger: 0.0009
Maintenance: 0.2953

Flight Crew: 0.2563
Passenger: 0.6599
Maintenance: 0.0837

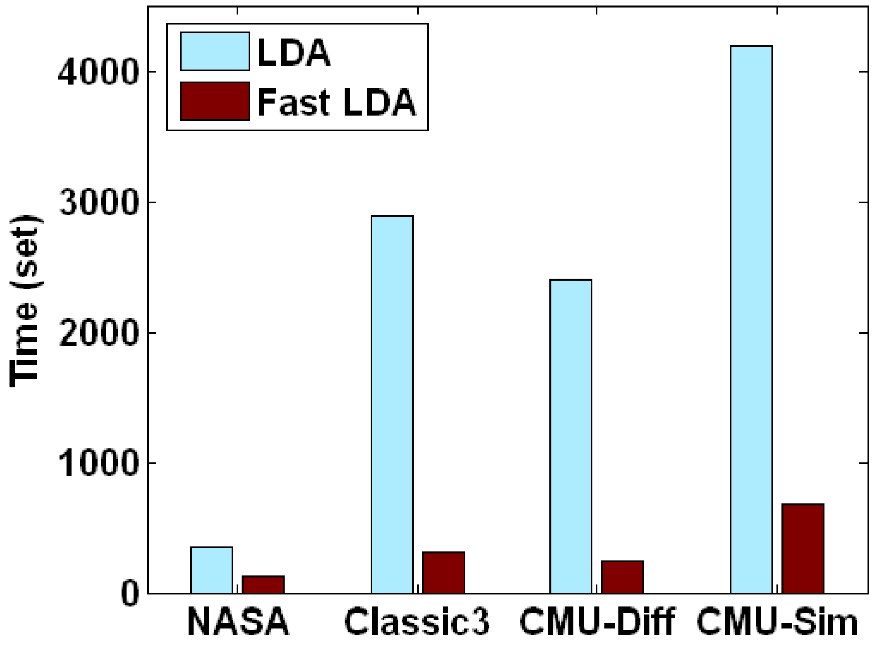
Flight Crew: 0.1405
Passenger: 0.0663
Maintenance: 0.7932

Flight Crew: 0.0013
Passenger: 0.0013
Maintenance: 0.9973

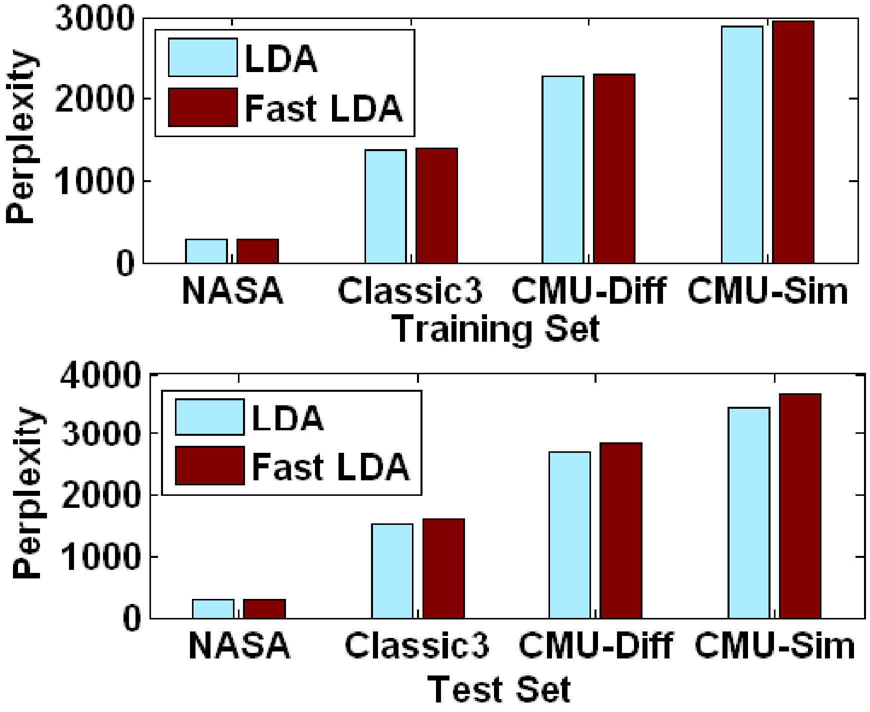
Red: Flight Crew Blue: Passenger Green: Maintenance

United Technologies Research Center

Fast LDA - A More Efficient Algorithm for LDA



Time comparison

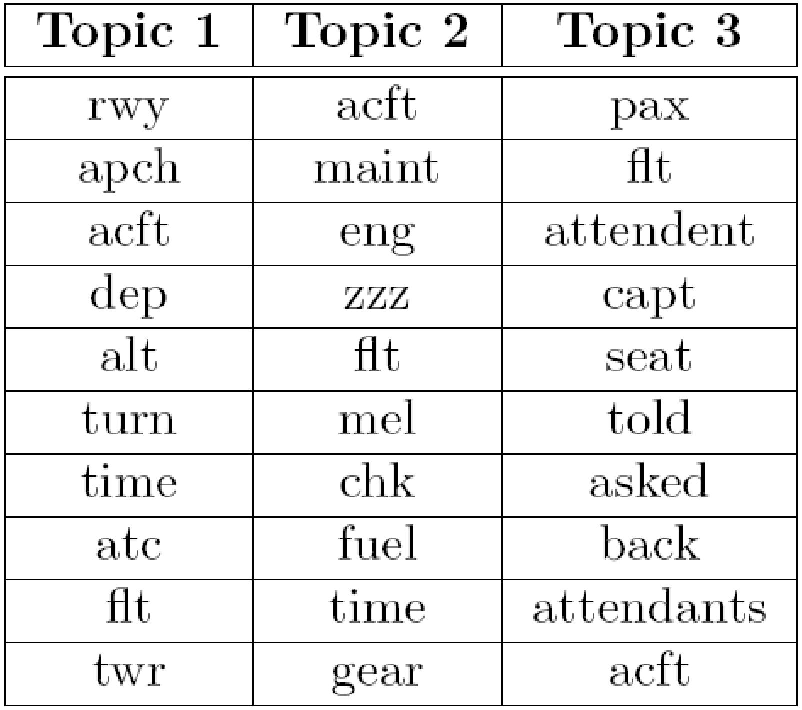


Perplexity comparison

- Perplexity is a monotonically decreasing function of log-likelihood, evaluating how the model fits the data –the lower the better.
- Fast LDA is much faster than LDA, with a similar perplexity.

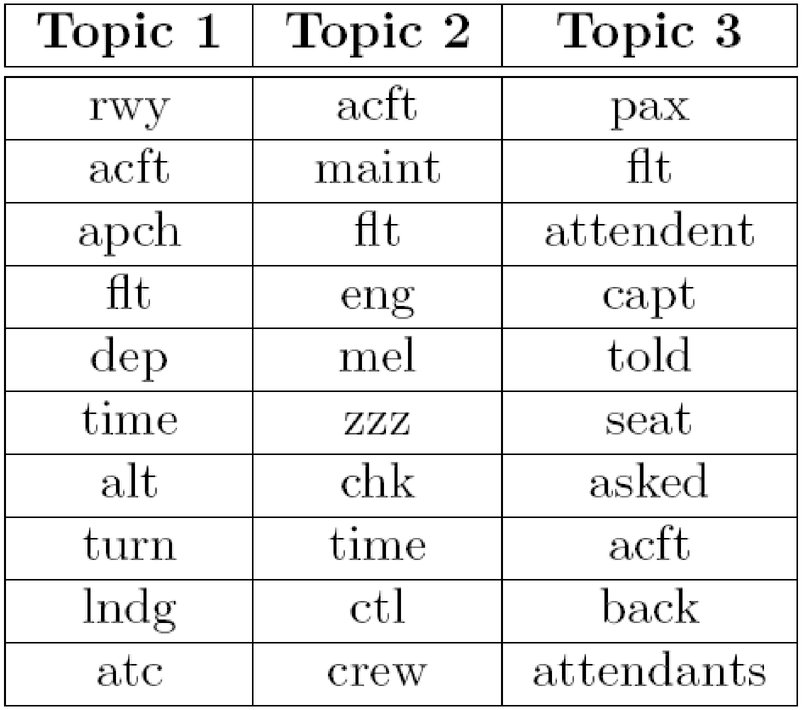
Word Lists for Topics

(a) LDA



Topic 1	Topic 2	Topic 3
rwy	acft	pax
apch	maint	flt
acft	eng	attendent
dep	zzz	capt
alt	flt	seat
turn	mel	told
time	chk	asked
atc	fuel	back
flt	time	attendants
twr	gear	acft

(b) Fast LDA



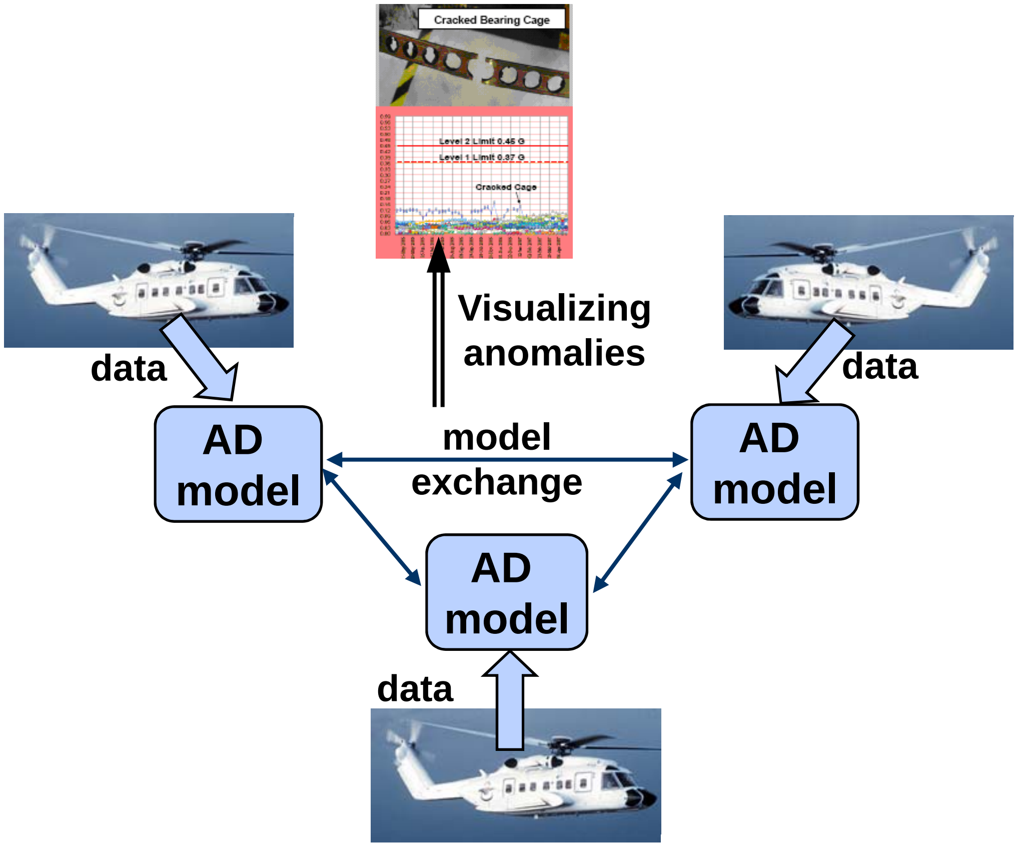
Topic 1	Topic 2	Topic 3
rwy	acft	pax
acft	maint	flt
apch	flt	attendent
flt	eng	capt
dep	mel	told
time	zzz	seat
alt	chk	asked
turn	time	acft
lndg	ctl	back
atc	crew	attendants

- Word lists from LDA and Fast LDA are similar.

Distributed Anomaly Detection

Objective of Research

Identify anomalous events or trends from multiple, homogeneous data sources



Visualizing anomalies

data

AD model

model exchange

AD model

data

AD model

data

Data Sources

- ADAPT System Data (obtained from NASA)
- Sikorsky S92 Flight Record Data
- Other publicly available non-aviation data sets

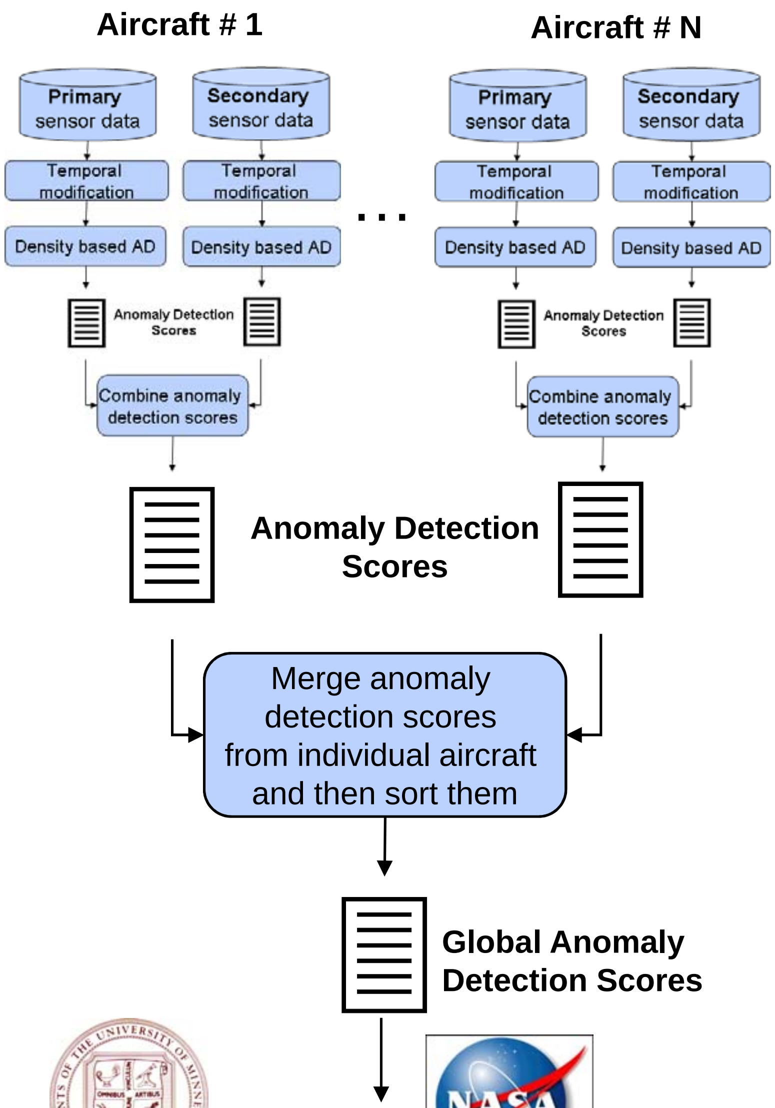
Key accomplishments:

- Evaluation of several types of anomaly detection algorithms
  - Density based methods (Parzen denisty estimator, local outlier factor)
  - Clustering based methods
  - Boundary based methods (unsupervised Support Vector Machines (SVM))
  - Reconstruction based methods (Minimal probability machine, auto-associative neural networks, Self-organizing maps (SOMs), minimum spanning trees)
- Development of several methods for anomaly detection from distributed sources:
  - Combining anomaly detection scores across distributed sites
  - Combining anomaly detection models among the distributed sites

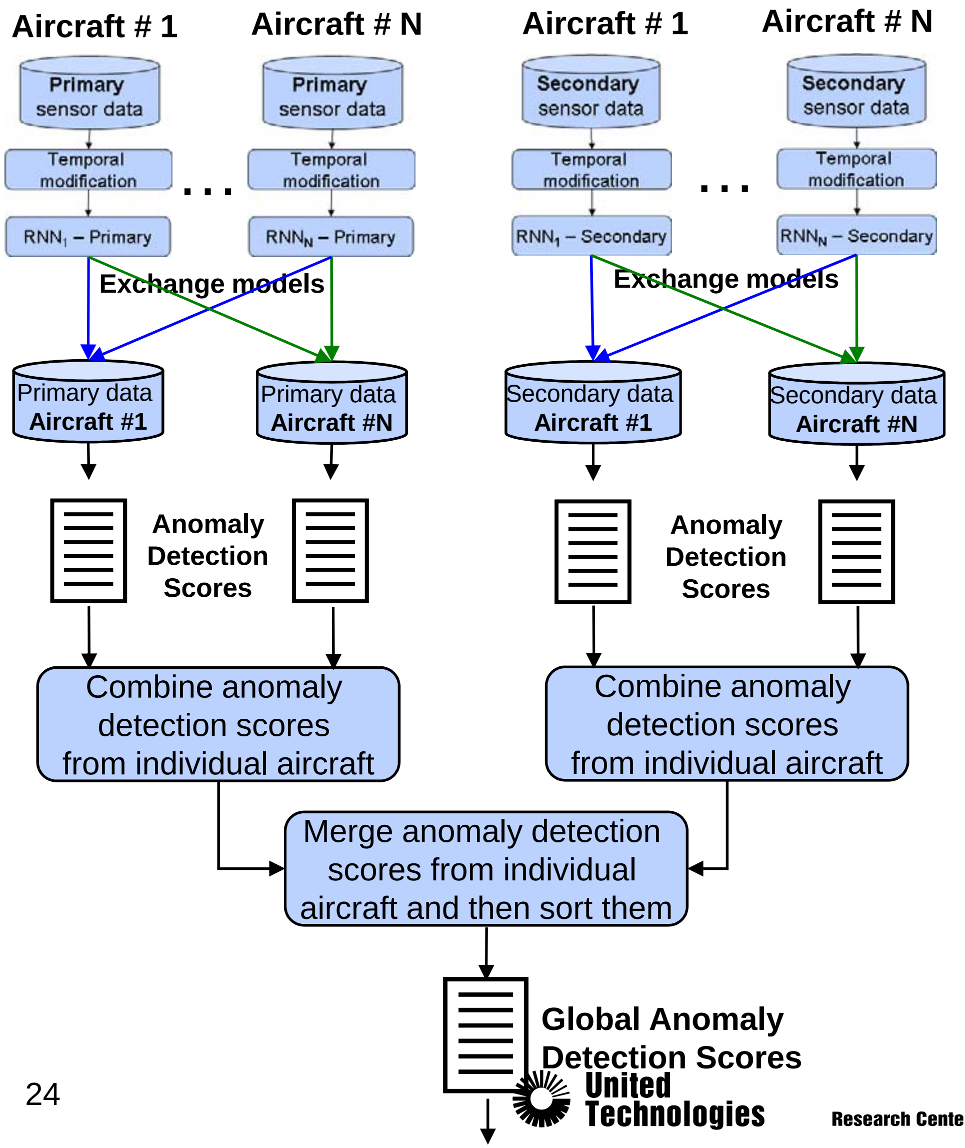
Combining Anomaly Detection (AD) Methods

Simple ranking and weighted voting

Combining Anomaly Detection Scores

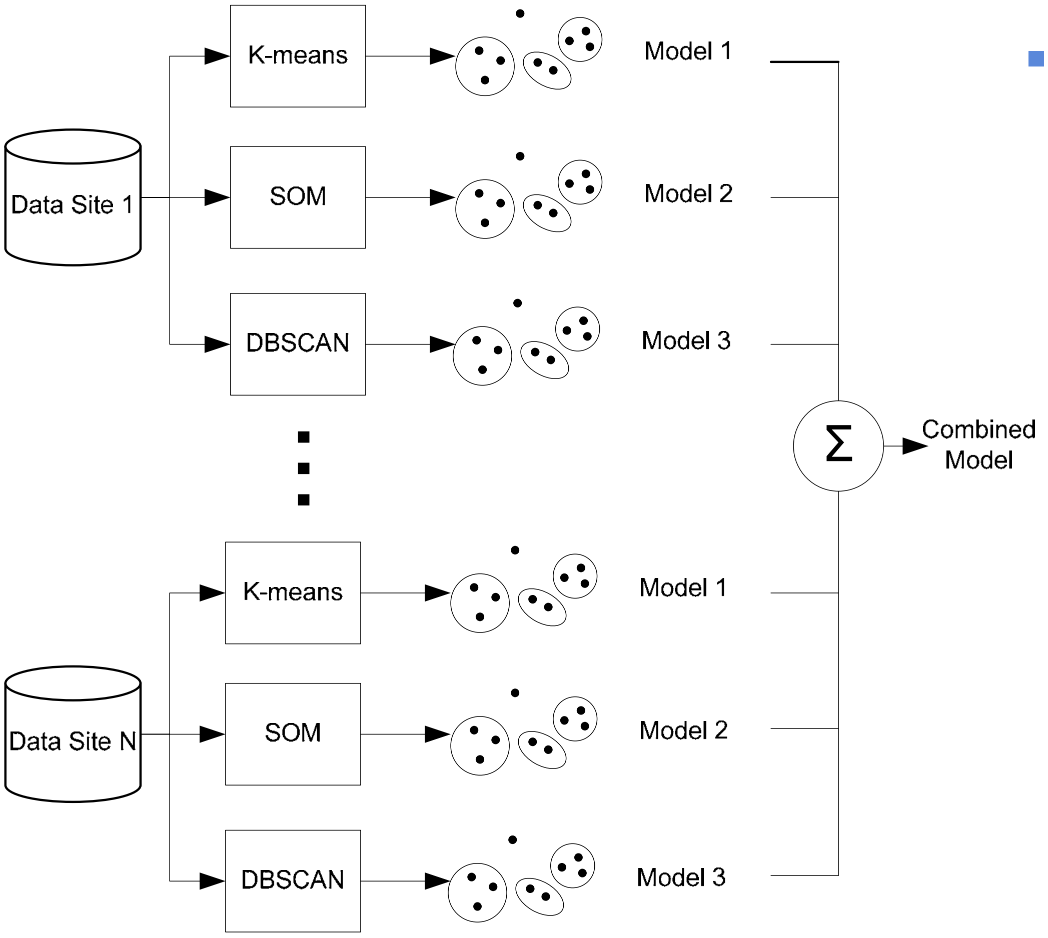


Combining Anomaly Detection Models



Combining Anomaly Detection (AD) Methods

Quality and diversity based combining



- Main idea:
  - Perform clustering and identify modes of normal behavior
  - Compute anomaly detection score as a Mahalanobis distance to the closest cluster
  - Build regression local models to learn anomaly detection score
  - Combine local modes to detect global anomalies by using both quality and diversity

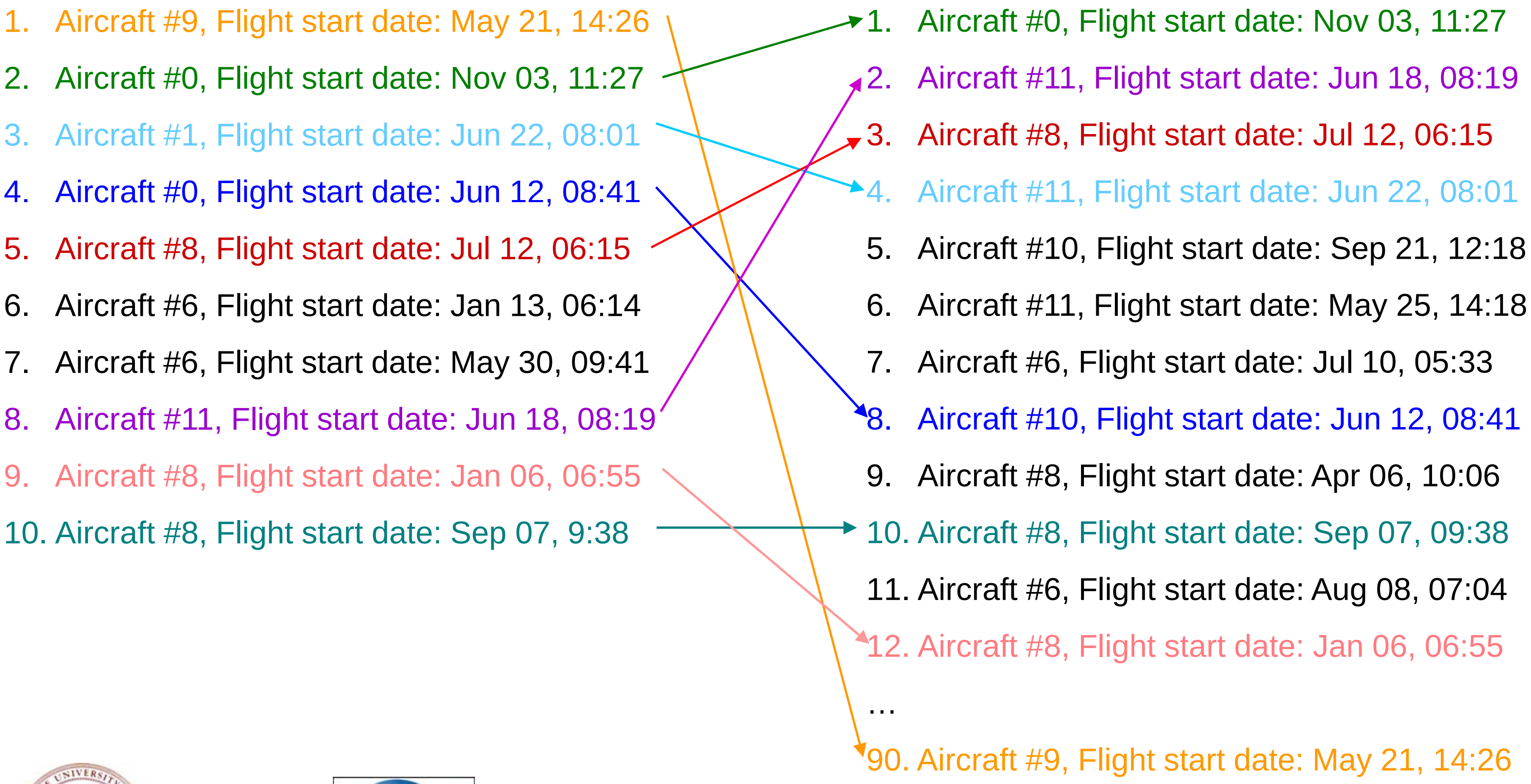
Methodology

- Combine local models' results by model quality and diversity
  - Quality - The performance of anomaly detection is related to the clustering quality of the uniform model
    - Silhouette index (SI) - reflecting the compactness and separation of clusters
    - Davies-Bouldin (DB) - Average similarity between each cluster
    - Dunn index (DI) - How similar the objects are within each cluster and how well the objects of different clusters are separated
    - Calinski-Harabasz (CH) - centroid intra-cluster and inter-cluster distances
  - Diversity- Diversity plays a significant role in combining prediction models, higher diversity leads to higher predict accuracy.
    - Adjusted Rand index (AR)
    - Jaccard index (JI)
    - Fowlkes-Mallows index (FM)

Combining Anomaly Detection models

Anomaly Detection on merged data from aircraft

Combining anomaly detection scores after applying AD algorithms on each individual aircraft



1. Aircraft #9, Flight start date: May 21, 14:26
2. Aircraft #0, Flight start date: Nov 03, 11:27
3. Aircraft #1, Flight start date: Jun 22, 08:01
4. Aircraft #0, Flight start date: Jun 12, 08:41
5. Aircraft #8, Flight start date: Jul 12, 06:15
6. Aircraft #6, Flight start date: Jan 13, 06:14
7. Aircraft #6, Flight start date: May 30, 09:41
8. Aircraft #11, Flight start date: Jun 18, 08:19
9. Aircraft #8, Flight start date: Jan 06, 06:55
10. Aircraft #8, Flight start date: Sep 07, 9:38

1. Aircraft #0, Flight start date: Nov 03, 11:27
2. Aircraft #11, Flight start date: Jun 18, 08:19
3. Aircraft #8, Flight start date: Jul 12, 06:15
4. Aircraft #11, Flight start date: Jun 22, 08:01
5. Aircraft #10, Flight start date: Sep 21, 12:18
6. Aircraft #11, Flight start date: May 25, 14:18
7. Aircraft #6, Flight start date: Jul 10, 05:33
8. Aircraft #10, Flight start date: Jun 12, 08:41
9. Aircraft #8, Flight start date: Apr 06, 10:06
10. Aircraft #8, Flight start date: Sep 07, 09:38
11. Aircraft #6, Flight start date: Aug 08, 07:04
12. Aircraft #8, Flight start date: Jan 06, 06:55

…

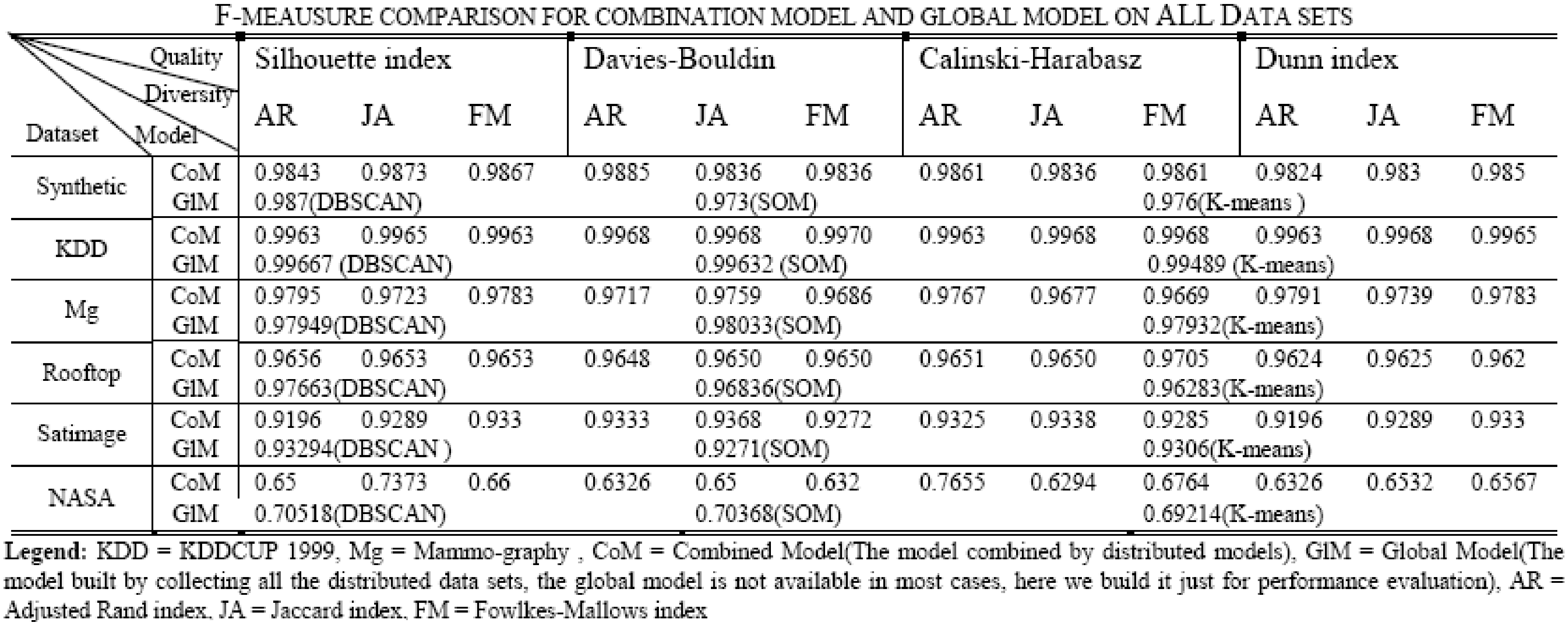
90. Aircraft #9, Flight start date: May 21, 14:26

United Technologies Research Center

Experiment results

- Set up
  - Data set:
    - Synthetic
    - KDDCUP 1999
    - Mammography
    - Rooftop
    - Satimage
    - NASA data
    - Sikorsky data
  - Data distributed into five (ten for KDD data) local sites
- Measures
  - F-value, Anomaly detection performance
  - Clustering quality, Local model quality
  - Agreement on test data, Local model diversity
  - Global model built by collected all local data sets, Comparison

Experiment results



F-MEAUSURE COMPARISON FOR COMBINATION MODEL AND GLOBAL MODEL ON ALL DATA SETS

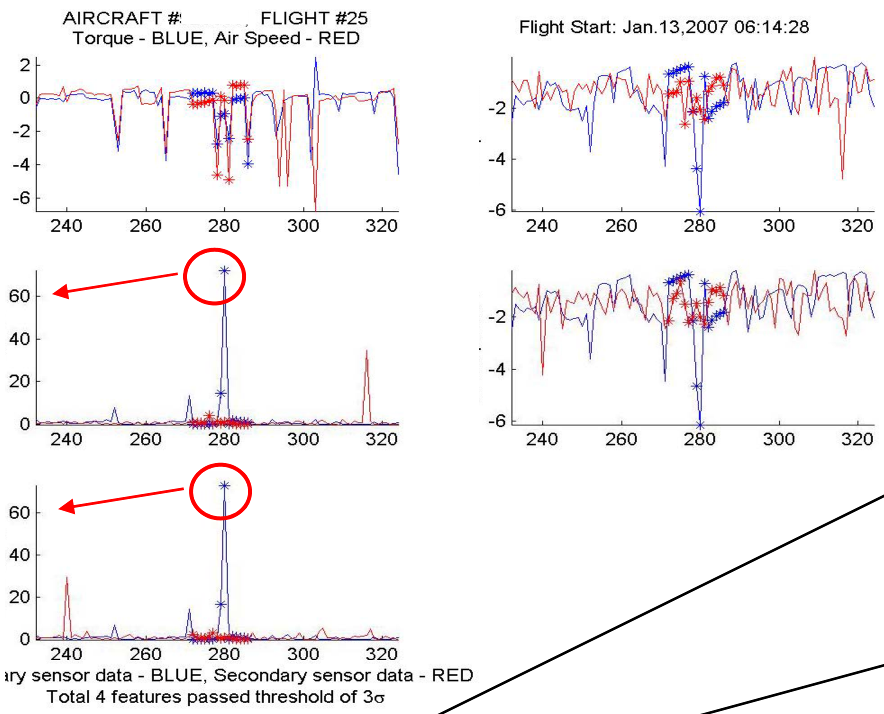
Quality / Diversity / Dataset / Model		Silhouette index			Davies-Bouldin			Calinski-Harabasz			Dunn index		
		AR	JA	FM	AR	JA	FM	AR	JA	FM	AR	JA	FM
Synthetic	CoM	0.9843	0.9873	0.9867	0.9885	0.9836	0.9836	0.9861	0.9836	0.9861	0.9824	0.983	0.985
	GlM	0.987(DBSCAN)				0.973(SOM)				0.976(K-means)			
KDD	CoM	0.9963	0.9965	0.9963	0.9968	0.9968	0.9970	0.9963	0.9968	0.9968	0.9963	0.9968	0.9965
	GlM	0.99667 (DBSCAN)				0.99632 (SOM)				0.99489 (K-means)			
Mg	CoM	0.9795	0.9723	0.9783	0.9717	0.9759	0.9686	0.9767	0.9677	0.9669	0.9791	0.9739	0.9783
	GlM	0.97949(DBSCAN)				0.98033(SOM)				0.97932(K-means)			
Rooftop	CoM	0.9656	0.9653	0.9653	0.9648	0.9650	0.9650	0.9651	0.9650	0.9705	0.9624	0.9625	0.962
	GlM	0.97663(DBSCAN)				0.96836(SOM)				0.96283(K-means)			
Satimage	CoM	0.9196	0.9289	0.933	0.9333	0.9368	0.9272	0.9325	0.9338	0.9285	0.9196	0.9289	0.933
	GlM	0.93294(DBSCAN)				0.9271(SOM)				0.9306(K-means)			
NASA	CoM	0.65	0.7373	0.66	0.6326	0.65	0.632	0.7655	0.6294	0.6764	0.6326	0.6532	0.6567
	GlM	0.70518(DBSCAN)				0.70368(SOM)				0.69214(K-means)			

Legend: KDD = KDDCUP 1999, Mg = Mammo-graphy , CoM = Combined Model(The model combined by distributed models), GlM = Global Model(The model built by collecting all the distributed data sets, the global model is not available in most cases, here we build it just for performance evaluation), AR = Adjusted Rand index, JA = Jaccard index, FM = Fowlkes-Mallows index

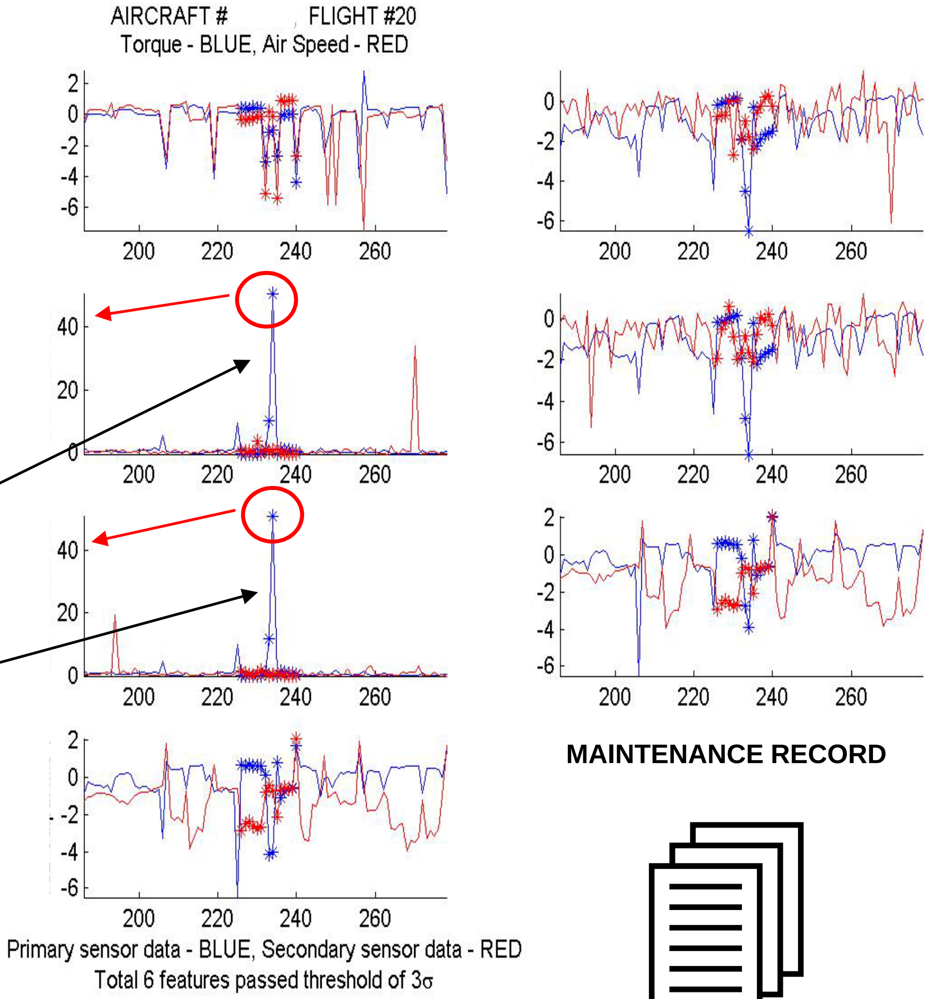
Combining Density-based Anomaly Detection Scores

Aircraft #Y, Flight start date: Jan 13, 06:14

AD results on merged data



Merging AD scores from individual aircraft



MAINTENANCE RECORD

These spikes are not that high (~40 σ far from 0 mean) as the spikes in the merged data (~60 σ far from 0 mean)

Publications

- Varun Chandola, Varun Mithal, Vipin Kumar, *Comparative Evaluation of Anomaly Detection Techniques for Sequence Data*, to appear in Proceedings of the IEEE Conference on Data Mining (ICDM), 2008.
- Varun Chandola, Arindam Banerjee, Vipin Kumar, *A Survey of Anomaly Detection*, to appear in ACM Computing Surveys, 2008.
- Hanhuai Shan, Arindam Banerjee, *Bayesian Co-Clustering*, 2008.
- William Schuler, Samir Abdel Rahman, Tim Miller, Lane Schwartz, *Robust Incremental Parsing using Human-Like Memory Constraints*, Journal of Computational Linguistics, 2008.
- Tim Miller, William Schuler, *An Empirical Evaluation of HHMM Parsing Time*, Proceedings of Midwest Computational Linguistics Conference, 2008.
- Junlin Zhou, Aleksander Lazarevic, Kyu-Wei Hsu, Nishith Pathak, Jaideep Srivastava, *Detecting Global Anomalies from Distributed Data Sources*, submitted to the Data Mining and Knowledge Discovery Journal, special issue on Outlier Analysis.
- Junlin Zhou, Aleksander Lazarevic, Kuo-Wei Hsu, Jaideep Srivastava, *Unsupervised Learning Based Distributed Detection of Global Anomalies*, submitted to SIAM Data Mining Conference, 2009.